DOCUMENT RESUME

ED 278 706 TM 870 133

AUTHOR Waller, Michael I.
TITLE Latent Trait Models and Random Guessing: A Comparison of Two Logistic Latent Trait Models.
PUB DATE Apr 86
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Tests; Cohort Analysis; Elementary Secondary Education; Equations (Mathematics); Estimation (Mathematics); *Goodness of Fit; *Guessing (Tests); *Item Analysis; *Latent Trait Theory; Mathematical Models; Maximum Likelihood Statistics; *Multiple Choice Tests; Testing Problems; Test Items
IDENTIFIERS Chi Square Test; *Three Parameter Model

ABSTRACT

This study compares the fit of the 3-parameter model to the Ability Removing Random Guessing (ARRG) model on data from a wide range of tests of cognitive ability in three representative samples. When the guessing parameters under the 3-parameter model are estimated individually for each item, the 3-parameter model yields the better fit to approximately 61 percent of these examinations. The ARRG model, which is designed to remove the effects of purely random guessing only, produces a better fit in 15 percent of these examinations, apparently to only those data sets which contain relatively little partial knowledge guessing behavior. The 2-parameter, free response model produces the best fit to 24 percent of these examinations. These results indicate that the 3-parameter model does make an adjustment for partial knowledge guessing. A comparison is also made between fits of the 3-parameter model with guessing parameters estimated at a constant value of one divided by the number of alternatives, and those estimated individually. Individually estimated guessing parameters are found to produce fits far superior to those estimated at a constant value. A solution to the convergence problems often obtained with the 3-parameter model is discussed. (Author/JAZ)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED278706

# Latent trait models and random guessing: A comparison of two logistic latent trait models.

Michael I. Waller

The University of Wisconsin-Milwaukee

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. I. Waller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

This paper was presented at the Annual Meeting of The American Educational Research Association, San Fransisco CA, April 19, 1986.

TM 870 133

## Abstract

This study compares the fit of the 3-parameter model to the ARRG model on data from a wide range of tests of cognitive ability in three representative samples. When the guessing parameters under the 3-parameter model are estimated individually for each item, the 3-parameter model yields the better fit to approximately 61% of these examinations. The ARRG model, which is designed to remove the effects of purely random guessing only, produces a better fit in 15% of these examinations, apparently to only those data sets which contain relatively little partial knowledge guessing behavior. The 2-parameter, free response model produces the best fit to 24% of these examinations. These results indicate that the 3-parameter model does make an adjustment for partial knowledge guessing.

A comparison is also made between fits of the 3-parameter model with guessing parameters estimated at a constant value of one divided by the number of alternatives, and those estimated individually. Individually estimated guessing parameters are found to produce fits far superior to those estimated at a constant value. A solution to the convergence problems often obtained with the 3-parameter model is discussed.

# Latent trait models and random guessing: A comparison of two logistic latent trait models.

Michael I. Waller

The University of Wisconsin-Milwaukee

## Introduction

Latent trait or item response theory (IRT) models were originally developed for use with free response items, but have come to be used extensively with multiple choice data. In such a context some revision of the model must be made in order to deal with the possibility of an examinee obtaining the correct answer independent of his or her knowledge and ability, but simply by means of a lucky random guess of the correct alternative. This study compares two revisions of the logistic form of the 2-parameter, free response latent trait model that deal with the problem of random guessing, the 3-parameter model (Birnbaum, 1968) and the Ability Removing Random Guessing (ARRG) model (Waller 1974). The models are compared in terms of their fit to the responses of three generally representative cohorts of children to eleven well known standardized examinations.

These two models are designed to adjust the 2-parameter, free response model to handle the effects of purely random guessing only, ignoring what has been called partial knowledge guessing, a type of test taking behavior which also may be assumed to exist in many testing situations. However one of the results of this study is that the 3-parameter model may be viewed as also making an

adjustment, at least in part, for partial knowledge guessing behavior.

## The Models

The free response logistic latent trait model is given by equation 1 where $P_{ij}$ represents the probability that person i gets item j correct given his or her ability:

$$P_{ij} = \frac{1}{1 + \exp( b_j - a_j \theta_i)}. \qquad (1)$$

Here $b_j$ and $a_j$ are item difficulty and discrimination respectively; and $\theta_i$ is person ability.

The most commonly used modification is that contained in Birnbaum's three-parameter model given in equation 2.

$$P_{ij} = g_j + (1 - g_j) * \frac{1}{1 + \exp( b_j - a_j \theta_i)} ; \qquad (2)$$

where $g_j$ represents the item guessing parameter or lower asymptote.

The three-parameter model has been found to have a number of problems. First, substantively this model can be viewed as postulating what appears to be an unreasonable situation for what happens when an item is encountered by a person. That is, this model has been interpreted as being one which assumes that people either guess totally randomly or answer on the basis of their knowledge. Under this interpretation the model is viewed as ignoring the often occurring situation of a person answering on the basis of partial knowledge; i.e., answering by eliminating one or more of the alternatives and choosing randomly from amongst the

remainder (Birnbaum, 1968).

Since one of the results of this study will be that this may be a limited view of the 3-parameter model, let us consider more fully the rational for this statement. If we let

$$F_{ij} = \frac{1}{1 + \exp( b_j - a_j \theta_i)}; \qquad (3)$$

then $F_{ij}$ can be viewed as representing the probability that examinee i knows the correct answer to item j; and the 3-parameter model of equation 2 may be written, after some algebraic manipulation, as in equation 4.

$$P_{ij} = F_{ij} * 1 + (1-F_{ij}) * g_j. \qquad (4)$$

We may view this equation as dividing the act of a person getting an item correct into two mutually exclusive events whose probabilities sum to give the probability of a correct response. The first term represents the probability of getting the item correct when the examinee knows the correct answer, while the second term represents this probability when the examinee knows nothing concerning the correct answer, and guesses randomly.

Both terms in equation 4 utilize the defining equation for conditional probability of event A given event B (Feller, 1968) for each of the two mutually exclusive events in the equation:

$$P(A,B) = P(B) * P(A \text{ given } B). \qquad (5)$$

That is, the right side of equation 4 may be interpreted as the probability that the examinee knows the correct answer ($F_{ij}$) multiplied by the probability that s/he will get the item correct

given that s/he knows the answer (1); plus, the probability that s/he doesn't know the correct answer ($1-F_{ij}$) multiplied by the probability that s/he will get the item correct, by guessing randomly, given that s/he doesn't know the correct answer ($g_j$). It is this view of the 3-parameter model that has been interpreted to mean that partial knowledge guessing is ignored by this model.

However, this view focuses on probabilities representing guessing behavior for those examinees in the lower portion of the range of abilities, in that portion where pure random guessing is most likely to occur. Let us consider the middle range of abilities for a particular item, focusing on the change in the probability of a correct response, $P_{ij}$, that will result from raising the lower asymptote. Figure 1 presents two ogives for an item of middle difficulty ($b_j = 0.0$) and discrimination ($a_j = 1.7$). The upper curve represents the 3-parameter model with guessing parameter simulating a four choice item ($g_j = .25$). The lower curve represents the free response, 2-parameter model for the same item.

---

Insert Figure 1 about here.

---

As we proceed along the curve, starting at the lowest abilities and proceeding to middle and higher abilities, we see that in comparison to the free response model of equation 2, $P_{ij}$ is raised in this portion of the curve as well. In fact, $P_{ij}$ is raised all along the curve with the change becoming smaller and smaller as we proceed from lower abilities through middle abilities to higher

abilities. It is only the asymptote that is constant representing the probability of a correct response from purely random guessing.

While the effect on $P_{ij}$ of raising the lower asymptote in this middle part of the ability continuum has not been the focus of much discussion, it is possible that the increase implied by the 3-parameter model in this range of the ability continuum could be an appropriate modeling of some of the effects of partial knowledge guessing. That is, it could be an appropriate, albeit approximate, upward correction in the modeling of the probability of a correct response in latent trait models to adjust for the act of eliminating one or more of the alternatives and choosing randomly from amongst the remainder.

As indicated below, the ARRG model may be used to isolate the effects of purely random guessing from other types of guessing behavior. Therefore this study will include an examination of the appropriateness of this adjustment made by the 3-parameter model to the free response, 2-parameter model by examining the fit of the 3-parameter model and comparing it to the fit of the ARRG model.

A second concern adherent to this model surrounds problems with convergence during estimation of the parameters of the model. Specifically, Samejima (1973) has shown that for some response vectors maximum likelihood estimates of ability are not unique and do not even exist at finite values. Simililar convergence problems have been found to exist during estimation of the item parameters. The results of this study yield some suggestions concerning

solutions to convergence problems of the item parameters.

The ARRG model, on the other hand, was designed to remove only the effects of purely random guessing in a manner that obviates convergence problems during estimation. This model builds on two characteristics inherent in all latent trait models; namely, that given a set of items that fit a latent trait model, one can obtain ability estimates on a common metric using any subset of such items; and that for each person and for each item to which that person responds one can routinely obtain an estimate of the probability of a correct response.

The latter characteristic allows us to identify those items which are very difficult for each person and therefore may attract purely random guessing; and the former characteristic allows us to eliminate such items from estimation of the person's ability while retaining the common metric for all resulting ability estimates even though different subsets of items will be used for different examinees. The ARRG model utilizes these two characteristics by assuming that when an examinee guesses in an essentially random manner s/he does so on only those items with a low probability of a correct response; that is, lower than the chance level of the item. Therefore the ARRG model estimates the probability of a correct response for each examinee's response to each item and ignores those responses for which the estimated probability of a correct response is quite low during estimation of that examinee's ability.

The model is characterized mathematically by the fact that it

is the 2-parameter, free response model presented in equation 1 modified to include only those interactions between an item and a person for which purely random guessing is unlikely to occur. Therefore this model can be characterized by equation 6.

$$P_{ij} = \frac{1}{1 + \exp( b_j - a_j \theta_i)}, \quad \text{for } P_{ij} > P_c; \qquad (6)$$

where $P_c < 1/A_j$; $A_j$ = the number of alternatives for item j.

The ARRG model essentially divides the item responses of <u>each</u> individual into two groups: those items for which $P_{ij} > P_c$, and those items for which $P_{ij} < P_c$; and then uses <u>only</u> those items with a reasonable probability of a correct response from factors other than purely random guessing, i.e. items from the first set, to estimate a person's ability. As has been shown (Waller 1974), the resulting elimination of some items under this model does not, as might be expected, reduce the precision of the resulting ability estimates. Rather the resulting precision increases because the noise resulting from the random responses has also been removed during estimation of the item parameters.

The ARRG model, then, individualizes the random guessing adjustment by determining a different set of items which is unique for each person upon which to estimate that person's ability. It has no convergence problems since once an individual's set of items has been selected, the model becomes the two-parameter model which produces stable parameter estimates if one assumes a prior distribution of abilitites as in the LOGOG computer program (Kolakowski and Bock, 1973).

The purpose of this study is to examine empirically the question of which of the two models produces the better fit to data in situ. This is accomplished by comparing the fit of these two models when each is applied to a large number of data sets measuring a wide variety of cognitive abilities from a nationally known and well established battery of various cognitive abilities.

Instrumentation and data

The measures used are the different content area examinations which comprise the Comprehensive Tests of Basic Skills (CTB McGraw Hill, 1974), Form S, levels 2,3 and 4. This battery of tests provides measures of the following cognitive abilities: reading vocabulary; reading comprehension; spelling; mechanics of language; language expression; mathematical computation; mathematical concepts; mathematical applications; reference skills; science; and social studies.

The examinees consist of samples from three cohorts of children obtained by stratified random sampling representative of the population of children in three grades from a southeastern state; each sample is of size just under 1000. The three cohorts of children are: elementary school(grade 4); middle school(grade 7); and high school(grade 10). Applying eleven content areas tc each of these cohorts yields a total of 33 different examinations upon which to compare the two models.

Method

The computer program used is an augmented version of LOGOG

(Kolakowski and Bock, 1973). In the procedure utilized by this program joint estimation of the two item parameters common to both models, difficulty and discrimination, is accomplished through conditional maximum likelihood assuming a particular distribution of abilities. A normal distribution of abilities was assumed for each ability examined in this study.

This assumption of a prior distribution of abilities made during estimation of the item parameters allows for estimation of all the parameters of the 2-parameter model without any problems of convergence. The prior assumption of a normal distribution of abilities is made throughout this study and may introduce some bias into the estimated values of these parameters in those cases where the distribution of abilities is sufficiently skewed. However, it seems reasonable to assume that the distribution of abilities in many of these cognitive abilities is nearly normal, and also, that any resulting bias in these estimated values would effect both models in essentially the same manner and can therefore be ignored in the comparisons made in this study.

For the three-parameter model the guessing parameter or lower asymptote for each item is estimated separately. Maximum likelihood estimation of this parameter has proven to be less than satisfactory; for example there is no mathematical lower bound on this parameter and joint maximum likelihood estimation of all three parameters of this model often yields estimates of the guessing parameter which are negative and consequently difficult to interpret. In addition the convergence problems referred to above

are magnified when this additional parameter is included in conditional maximum likelihood estimation.

There are at least two methods for estimating this parameter currently in use. The simplest method and perhaps the method used most often might be called the constant method. In the constant method the guessing parameter is set to the same value for all items, 1/A; where A is the number of alternatives. The second method might be called the individualized method. In this method the value of $g_j$ for each item is obtained by ranking the examinees, distributing them into some number of groups or fractiles, and then estimating the lower asymptote graphically by examining the proportion correct in each of these fractiles. The individualized method with ten fractiles was used in this study for the comparisons between the two models. A second set of analyses utilizing the constant method was also performed, and the resulting fits compared to the 3-parameter model utilizing individualized method and to the ARRG model.

For the ARRG model one $P_c$ value is used for each content area examination. The estimation procedure for this parameter is a gross minimum Chi Square procedure. Values for $P_c$ are chosen from 0 to 1/A in steps of .05; e.g., for an examination consisting of four choice items such as the CTBS the trial values would be .00, .05, .10, .15, .20, and .25. For each value of $P_c$ a complete item analysis and test of fit is performed. The final estimate of $P_c$ is that value yielding the minimum Chi Square value utilized to test

the fit of the instrument as a whole.

The test statistic utilized is the standard Pearsonian Chi Square goodness-of-fit statistic. The cells required for this statistic are formed as described above concerning estimation of the item parameters by ranking the examinees by ability and then dividing the sample into ten fractiles; the number of examinees in each group is again chosen to reflect a normal distribution of abilities. The assumption of normality does not effect the validity of this test statistic.

One comparison between the two models is made for each content area examination for each of the three cohorts. The number of degrees of freedom utilized to test the fit of each item of each examination is equal to the number of cells (10) minus the number of parameters estimated, yielding 7 for the 3-parameter model and 8 for the ARRG model. The value of the Pearsonian Chi Square for each item is summed to get a value of Chi Square which may be used for testing the fit of each of the two models for each content area examination as a whole. The number of degrees of freedom for this test is the sum of the degrees of freedom for each item minus two. These two additional degrees of freedom are lost since the mean and standard deviation of the distribution of abilities are assigned values of zero and one respectively (Kolakowski and Bock, 1973). For the ARRG procedure an additional degree of freedom is also subtracted when making the test of fit of each instrument as a whole since this procedure requires estimation of one cutoff value, $P_c$, for use with all items in each examination.

Results and Discussion

The results of the 33 comparisons between the two models are presented in Table 1 in terms of the ratio of Chi Square to the degrees of freedom. These results indicate that in 61% of these examinations the 3-parameter model with individually estimated guessing parameters produced the best fit to the data, 20 out of 33. The ARRG model produced the best fit in five of the 33 possible examinations; while a 2-parameter or free response analysis produced the best fit for eight examinations.

---

Insert Table 1 about here.

---

These results indicate that the 3-parameter model with individually estimated guessing parameters was found to be generally superior to the ARRG model in the important area of producing the best fit to empirical responses to standardized tests developed via classical test theory. There are, of course, other possible criteria. One could, for example, choose the model which requires elimination of the fewest items in order to produce an acceptable fit to such a test. However, there is no reason to expect that the results of such a study would be significantly different from the results of this study.

There were five examinations in which the ARRG model produced the best fit. Substantive explanations indicating the type of examination in which the ARRG model will produce the better fit did not appear to exist for any of these examinations. For example,

content analyses of the items in these examinations did not reveal any obvious difference between these examinations and those in which the 3-parameter model with individually estimated guessing parameters produced the best fit.

In terms of methodological considerations it was the case that in four of the five examinations in which the ARRG model produced the best fit, the best fitting cutoff probability, $P_c$, was equal to .05. This might suggest that the ARRG model may produce the superior fit to data in which random guessing occurs only on very difficult items. However in six other examinations the best fitting ARRG analysis was also at the $P_c$ = .05 level and still the 3-parameter model with individually estimated guessing parameters yielded the best fit to the data.

Within this context consider the fact that in ten of the analyses in which the 3-parameter model produced the best fit, the best fitting ARRG analysis was at the $P_c$ = .00 level. ARRG analyses characterized by $P_c$ =.00 producing the best fit may be interpreted as indicating little or no purely random guessing in the data; and yet the improvement in fit achieved by the 3-parameter model in these instances was usually substantial.

If we now focus on the fact that middle level ability examinees' probability of a correct response is also raised for a particular item when utilizing the 3-parameter model, the fact that the 3-parameter model produced a better fit in situations in which the ARRG model indicated little or no purely random guessing

present in the data may imply not only that the 3-parameter model represents a more accurate procedure than the ARRG model for handling the effects of partial knowledge guessing. More importantly, perhaps, these comparisons appear to imply that the 3-parameter model does not ignore partial knowledge guesing behavior entirely; rather, the 3-parameter model does make an adjustment, if only a partial one, for such guessing behavior.

These considerations would also indicate a possible explanation as to why there existed examinations in which the ARRG model produced a fit superior to the 3-parameter model. Such results suggest that the ARRG model will produce a better fit than the 3-parameter model in situations in which there is little or no partial knowledge guessing, but in which there exists a significant amount of purely random guessing.

A further examination of Table 1 reveals that the best fit was produced by the 2-parameter, free response model for a significant proportion of all the examinations studied here (24%). Therefore, one conclusion which may be drawn from this study would appear to be that both an ARRG analysis, which implicitly includes a free response analysis when $P_c = 0.00$, and a 3-parameter analysis should be performed and the best fitting model chosen from among all the analyses.

A second set of results yielded by this study is a comparison of the fit of the 3-parameter model utilizing the constant method for estimating the guessing parameter of all items as opposed to

the 3-parameter method utilizing guessing parameters estimated by the individualized method. The results are presented in Table 2 in terms of the ratio of Chi Square to degrees of freedom.

---

Insert Table 2 about here.

---

In the analyses under the 3-parameter model utilizing the constant method, the appropriate number of degrees of freedom to use for testing the examination as a whole is not obvious. It can be argued that the number of degrees of freedom should be the same as that described above for testing the fit of any 3-parameter model: Number of groups minus the number of parameters estimated, three, multiplied by the number of items. Within the structure of the LOGOG program this would yield 7 multiplied by the number of items. However it could also be argued that since only one value is used for all items, the correct number of degrees of freedom for testing the fit of the instrument as a whole is: Number of groups minus the number of parameters fit to each item, two, multiplied by the number of items with only one subtracted from this total for the one value used for all the guessing parameters.

The latter degrees of freedom formula is more than the former formula by a value equal to the number of items minus one; and therefore places the constant estimation method in it's best light. Since the results of these comparisons utilizing either formula indicate that individually fit guessing parameters produce the best fit, the number of degrees of freedom used as presented in Table 2 for testing the fit of the model when a constant value is used for

the estimating guessing parameter is the latter formula.

The results of these analyses are that in 82%, or 27 out of 33 examinations, the 3-parameter model utilizing the individualized method produced a better fit to the data than did the 3-parameter model utilizing the constant method. Comparing Table 2 to Table 1 one observes that the ARRG model also produced a better fit to these data than the 3-parameter model utilizing the constant method in 75% of 32 of these examinations (in one examination both produced the same fit). It should also be noted that these results do not effect, in any way, the results of the comparison between the 3-parameter model and the ARRG model. Rather, the comparisons between the ARRG model and the 3-parameter model utilizing the constant method reinforce the contention that the guessing parameters should be estimated individually rather than at a constant value.

There were, as might be expected, convergence problems with the 3-parameter model in a number of testing situations. These problems occurred, with one exception, with items which were at least moderately difficult, indicating some guessing may be present in the data, and with items which at the same time had low overall discriminations. Recently possible mathematical solutions to convergence problems in latent trait models have been presented (e.g. Swaminathan and Gifford, 1985); however the present study indicates that such endeavors are not necessarily the best direction for future efforts.

For each of the problem items in this study convergence was obtained simply by setting the guessing parameter back to zero, thus fitting a 2-parameter model, and allowing the discrimination for the effected item to remain relatively low. Swaminathan and Gifford's work imply that the convergence problems in the 3-parameter model may be solved by assuming an a priori distribution on the discriminations. That is, if one assumes a distribution of abilities and a distribution of discriminations, all three item parameters would converge. As Swaminathan and Gifford indicate, however, there is little substantive reason upon which to base a choice for a prior distribution of discriminations.

While "fixes" such as these are mathematically sound, it can be argued that they ignore the fact that the real world doesn't, at times, act as though these unidimensional models adequately discribe the testing situation (see, e.g., Ansley and Forsyth, 1985). The lack of convergence in such situations is not necessarily the result of estimation problems, but is more likely the result of the data telling us that our unidimensional latent trait models are at times quite limited. It may be that we must be ready to accept the reduced information provided by such items which will be reflected in lower discriminations as long as we insist on describing certain aspects of the world with one number when in fact it requires more than one.

In this study as in others (Hambleton and Murray, 1983; Traub, 1983; Loyd and Hoover, 1980), an adequate fit to the data was usually not obtained. (The one exception to this in these data sets

was the Language Expression examination in the Middle School cohort.) Such lack of fit is typical of latent trait analyses of standardized tests constructed via the classical test theory model. It is argued here, mathematically and substantively, that this is necessarily a result of the fact that the unidimensional assumption made by latent trait models is necessarily violated in these cognitive testing situations.

Mathematically the argument proceeds along the following lines: There are only two assumptions made by latent trait models: that the form of the response relation is as stated by the defining equation, the logistic in this case; and that the underlying latent trait is unidimensional (Lord and Novick, 1968). If the ability in question follows a normal distribution, Lord and Novick also show that the correct response relation is the normal response relation; and since the logistic response relation equation uniformly differs from the normal response relation equation by less than .01, no practical differences in tests of fit will be discerned between these two models. We may therefore conclude that where the distribution of abilities is approximately normal, the only source of lack of fit is the assumption of one underlying dimension. The problem of grouping adjacent ability levels when testing fit analagous to situations requiring Sheppard's correction (Kendall and Stuart, 1977) can be expected to have only a minor effect when compared to the large amount of lack of fit usually observed with these models; however it is likely that a Sheppard type correction would improve the fit of these models somewhat.

Substantively, consider tests that have been constructed to fit one of our unidimensional latent trait models. It is almost always the case that during instrument construction of such examinations items must be removed do to a lack fit which subject matter experts judge should be retained in order to test a particular ability in that area of cognitive functioning. Such ill-fitting items are almost always found to be testing an ability that is known by the subject matter experts to be multidimensional.

In summary, these comparisons of two logistic latent trait models designed to handle the problem of random guessing indicate the following: that the ARRG model produces the best fit to a small percentage of data sets, apparently those containing only purely random guessing but no significant partial knowledge guessing; that the 3-parameter model utilizing the individualized method for estimating the guessing parameter produces the best fit to the predominantly occurring situation of data containing both purely random guessing and significant amounts of partial knowledge guessing; that a significant proportion of data sets are best fit by a free response model; and that, consequently, both an ARRG analysis which implicitly includes a free response analysis and a 3-parameter analysis should be performed and the best fitting analysis chosen from among them.

The results also contain two other important implications for the practical use of latent trait models: First, that the guessing parameter in the 3-parameter model should be estimated individually and not with one value equal to 1/A. Second, utilizing the Bock

procedure as implemented in LOGOG of assuming a prior distribution of abilities, problems of convergence with items containing more than one dimension can regularly be solved by simply resetting the guessing parameter for those items to zero, and accepting the resulting lower discrimination as a necessary result when substantive considerations require employing a unidimensional model with multidimensional data.

References

Ansley, T.N. and Forsyth, R.A.(1985). An examination of the unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, Vol.9(1), pp.37-48.

Birnbaum, A.(1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M. Statistical Theories of MentalTest Scores. Reading MA: Addison Wesley.

CTB McGraw Hill (1974). Comprehensive Tests of Basic Skills. Monterey, CA.: McGraw Hill Inc.

Feller, William (1968). An Introduction to Probability Theory and It's Applications, Volume 1. NY: Wiley.

Hambleton, R.K. and Murray, L. (1983). Some goodness of fit investigations for item response models. In Hambleton, R.A.(Ed.), Applications of Item Response Theory. British Columbia: Educational REsearch Institute of British Columbia.

Kolakowski, D. and Bock, R.D. (1973). LOGOG: Maximum Likelihood Item Analysis and Test Scoring: Logistic Model For Multiple Item Responses. Ann Arbor MI: National Educational Resources.

Kendall, M.G. and Stuart, A. (1977). The Advanced Theory of Statistics, Volume 1, 4th ed.. NY: Hafner.

Lord, F.M. and Novick, M. (1968). Statistical Theories of Mental Test Scores. Reading MA.: Addison Wesley.

Loyd, B.H. and Hoover, H.D. (1980). Vertical equating using the Rasch Model. Journal of Educational Measurement Vol. 17, pp.178-193.

Samejima, Fumiko (1973) A comment on Birnbaum's three parameter model in the latent trait theory. Psychometrika Vol. 38, pp.221-233.

Swaminathan, H. and Gifford, J.A. (1985). Baysian estimation in the two-parameter logistic model. Psychometrika Vol. 50(3), pp.349-364.

Traub, R.E. (1983). Apriori considerations in choosing an item response model. In Hambleton, R.K.(Ed.) Applications of Item Response Theory. British Columbia: Educational Research Institute o itish Columbia.

Waller, Michael (1974). Removing the Effects of Random Guessing from ent Trait Ability Estimates. RB-74-32, Princeton NJ.: Educational Testing Service.

Table 1

Goodness-of-Fit Ratios
Chi Sq./ D.F. - 3-Parameter and ARRG by Cohort

| Content Area | Elementary (4th) 3-Parm | | ARRG | | Middle (7th) 3-Parm | | ARRG | | High (10th) 3-Parm | | ARRG | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reading Vocabulary | 2.03 | * | 2.58 | | 1.56 | * | 1.89 | | 1.67 | * | 2.41 | |
| | (40 items) | | | | (40 items) | | | | (40 items) | | | |
| Reading Comprehension | 2.33 | * | 2.60 | | 1.63 | | 1.60 | * | 1.80 | | 1.70 | F |
| | (45 items) | | | | (45 items) | | | | (45 items) | | | |
| Language, Spelling | 1.86 | * | 1.87 | | 1.25 | * | 1.56 | | 1.21 | * | 1.40 | |
| | (50 tems) | | | | (30 items) | | | | (30 items) | | | |
| Language Mechanics | 1.90 | | 1.63 | * | 1.39 | | 1.39 | F | 2.07 | | 1.71 | F |
| | (20 items) | | | | (20 items) | | | | (20 items) | | | |
| Language Expression | 1.88 | * | 2.08 | | 1.12 | * | 1.28 | | 2.15 | | 1.81 | * |
| | (35 items) | | | | (35 items) | | | | (35 items) | | | |
| Mathmeatics Computation | 1.59 | * | 1.65 | | 2.14 | * | 2.63 | | 1.48 | * | 1.68 | |
| | (48 items) | | | | (48 items) | | | | (48 items) | | | |
| Mathematics Concepts | 2.90 | | 2.90 | F | 1.51 | | 1.43 | F | 1.48 | * | 1.51 | |
| | (25 items) | | | | (25 items) | | | | (25 items) | | | |
| Mathematics Application | 1.59 | * | 1.63 | | 1.81 | * | 2.44 | | 1.63 | * | 1.88 | |
| | (25 items) | | | | (25 items) | | | | (25 items) | | | |
| Reference Skills | 1.89 | | 1.75 | * | 1.67 | | 1.50 | F | 2.22 | | 1.93 | F |
| | (20 items) | | | | (20 items) | | | | (20 items) | | | |
| Science | 2.02 | | 1.90 | F | 1.76 | | 1.70 | * | 1.42 | * | 1.61 | |
| | (36 items) | | | | (41 items) | | | | (40 items) | | | |
| Social Studies | 1.59 | * | 1.88 | | 1.71 | * | 1.87 | | 1.58 | * | 1.81 | |
| | (37 items) | | | | (40 items) | | | | (39 items) | | | |

* indicates the model with the better fit.
F indicates free response produced the best fit.

Table 2

Goodness-of-Fit Ratios
Chi Sq./ D.F. - 3-Parameter Model
Individualized and Fixed Guessing Parameters

Cohort

| Content Area | Elementary (4th) Indvl | Elementary (4th) Fixed | Middle (7th) Indvl | Middle (7th) Fixed | High (10th) Indvl | High (10th) Fixed |
|---|---|---|---|---|---|---|
| Reading Vocabulary | 2.03 * (40 items) | 2.29 | 1.56 * (40 items) | 1.89 | 1.67 * (40 items) | 2.04 |
| Reading Comprehension | 2.33 (45 items) | 2.17 * | 1.63 * (45 items) | 3.02 | 1.80 * (45 items) | 1.89 |
| Language, Spelling | 1.86 (50 tems) | 1.79 * | 1.25 (30 items) | 1.23 * | 1.21 * (30 items) | 6.99 |
| Language Mechanics | 1.90 * (20 items) | 1.95 | 1.39 * (20 items) | 1.88 | 2.07 (20 items) | 1.87 * |
| Language Expression | 1.88 * (35 items) | 2.14 | 1.12 * (35 items) | 2.01 | 2.15 (35 items) | 2.05 * |
| Mathmeatics Computation | 1.59 * (48 items) | 1.83 | 2.14 * (48 items) | 2.58 | 1.48 * (48 items) | 2.28 |
| Mathematics Concepts | 2.90 * (25 items) | 3.79 | 1.51 * (25 items) | 3.40 | 1.48 * (25 items) | 1.79 |
| Mathematics Application | 1.59 (25 items) | 1.48 * | 1.81 * (25 items) | 2.12 | 1.63 * (25 items) | 2.84 |
| Reference Skills | 1.89 * (20 items) | 2.54 | 1.67 * (20 items) | 1.69 | 2.22 * (20 items) | 7.45 |
| Science | 2.02 * (36 items) | 2.80 | 1.76 * (41 items) | 2.98 | 1.42 * (40 items) | 1.59 |
| Social Studies | 1.59 * (37 items) | 2.31 | 1.71 * (40 items) | 4.98 | 1.58 * (39 items) | 5.33 |

* indicates the better fit.

**Figure 1**

**2-Parameter and 3-Parameter Ogives for a 4 Choice Item of Medium Difficulty**